ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 457 - 461 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

EFFECTIVENESS AND ROBUSTNESS OF HETEROGENEOUS WEB DOCUMENTS

MAMATHA GUNDI *,mamathagundi88@gmil.com

Md. Asim#, asim.cse36@gmail.com

ABSTRACT:

Now-a-days Internet [1][2] plays major role in our daily life such as e-commerce, e-seva etc. Usage of internet is increasing drastically more and more. World Wide Web (WWW) is widely used to publish and access information on the Internet. The web pages in many websites are automatically populated by using common templates with contents. It increase the bottleneck on the end users system. It reduce the system performance and speed of the operation. To reduce bottleneck on the web pages, a novel approach was presented from heterogeneous web documents. Represent the documents and path using matrix and it uses the MDL principle to manage the unknown no. of clusters. The Min Hash technique [6] to speed up the clustering process. It reduce the cost and maintenance.

Key words: Internet, Min Hash Technique, web pages, MDL, SDLC.

I. INTRODUCTION

For human beings, the templates provide readers easy access to the contents guided by consistent structures even though the templates are not explicitly announced. The unknown templates are considered harmful, it degrade the accuracy and performance. Thus, template detection and extraction techniques have received a lot of attention recently to improve the performance of web applications.

The problem of extracting a template from the web documents conforming to a common template has been studied in. However, in real applications, it is not trivial to classify. Since subtle changes in scripts or CGI parameters may result in a significant difference, we cannot simply group the web documents by URL and apply these methods for each group separately. If we use only URLs to group pages and it included in the same cluster.

II. SYSTEM REQUIREMENTS:

It includes a set of use cases that describe all the interactions the users will have with the software. The SRS also contains non-functional requirements. They impose constraints on the design or implementation.

i)System requirements specification:

A structured collection of information that embodies the requirements of a system. Within the SDLC[3] domain, the BA typically performs a liaison function between the business side of an enterprise and the information technology department or external service providers. Projects are subject to three sorts of requirements:Business requirements, Product requirements, Process requirements.

For example, a maximum process requirement may be imposed to help achieve a maximum sales price requirement ; a requirement that the product be maintainable often is addressed by imposing requirements to follow particular development styles. Requirements are either functional or non functional.

Functional Requirement. It specify something that the delivered system must be able to do. It include usability, maintainability etc. A collection of requirements define the characteristics or features of the desired system.

ii) Functional Requirements :

1) Create New User Registration

2) Admin Login

3) Upload the data information

4) View the User Information

5) The WebPages in many websites are automatically populated by using the common templates with contents.

6) The templates provide readers easy access to the contents guided by consistent structures.

7) We present novel algorithms for extracting templates from a large number of web documents

8) We cluster the web documents based on the similarity of underlying template structures in the documents so that the template for each cluster is extracted simultaneously.

9) We develop a novel goodness measure with its fast approximation for clustering and provide comprehensive analysis of our algorithm.

10) Our experimental results with real-life data sets confirm the effectiveness and robustness of our algorithm compared

iii) Non Functional Requirements :

The System non-functional Requirements are Usability, Reliability, Performance, Supportability, Implementation.

III. SYSTEM ANALYSIS

SDLC[4] is the process of creating or altering systems, and the models and methodologies. SDLC concept underpins many kinds of software development methodologies. It form the framework for planning and controlling the creation of an information system the software development process.

i.) **Existing System:**

An HTML document can be naturally represented with a DOM tree[8], web documents are considered as trees and many existing similarity measures for trees have been investigated for clustering. Disadvantage is, it is very expensive with tree-related distance measures. Clustering on sampled web documents is used to practically handle a large number of web documents.
The problem of extracting a template from the web documents conforming to a common template has been studied in. The solutions for this problem are applicable only when all documents are guaranteed to conform to a common template. We cannot classify massively crawled documents into homogeneous partitions in order to use these techniques

ii) Proposed System:

The page-level template detection where the template is computed within a single document. There are different systems proposed by many persons like Lerman et al., Zhai and Liu, Chakrabarti et al.. Our algorithms to be presented later represent web documents as a matrix and find clusters with the matrix. Biclustering[7] or co-clustering techniques are deal with a matrix . Coclustering algorithms find simultaneous clustering of the rows and columns of a matrix and require the numbers of clusters of columns and rows. Cluster only documents not paths. and the no. of clusters of columns and rows are unknown.

We propose to represent a web document and a template as a set of paths in a DOM tree. It uses XML query language and XPATH [2], paths are sufficient to express tree structures. By considering only paths, the overhead to measure the similarity between documents becomes small without significant loss of information.

IV. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. It is the most critical stage in achieving a successful new system and be effective. The implementation constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Main Modules

1. HTML Documents And Document Object Model.

2. The DOM defines a standard for accessing documents.

3. The DOM presents an HTML documents as a tree structure.

4. The entire documents is a document node, every HTML element is a element node, the texts in the HTML element is the text nodes ,every HTML attribute is an attribute node and comment are the comment nodes.

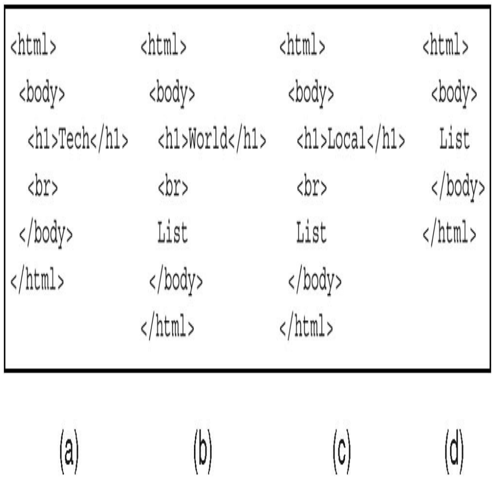
i)**Essential paths and templates**

1. Essential or useful paths have been selected and then *support* of the path is defined.

2. Then we provide *threshold* for paths in the document. According to threshold value the path is essential or not is identified.

3. If the *suport* of the path is **greater than or equal to** the *threshold value* the path is identified as the essential path .

4. Also we find out the ME i.e Matrix with essential paths.

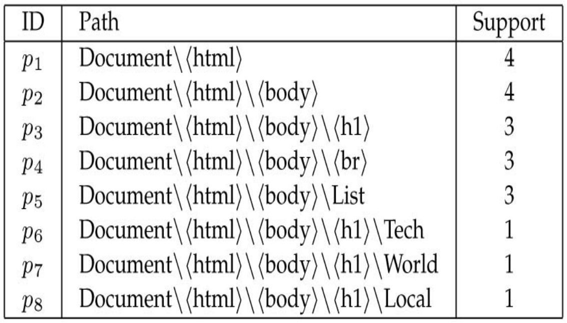


```
<html>          <html>            <html>            <html>
 <body>          <body>            <body>            <body>
  <h1>Tech</h1>   <h1>World</h1>    <h1>Local</h1>    List
  <br>            <br>              <br>             </body>
 </body>          List              List            </html>
</html>          </body>           </body>
                </html>           </html>

   (a)             (b)               (c)              (d)
```

Figh1 Simple web documents (a) Document d_1 , (b) Document d_2 , (c) Document $d_{3,}$

(d) Document d_4



ID	Path	Support
p_1	Document\⟨html⟩	4
p_2	Document\⟨html⟩\⟨body⟩	4
p_3	Document\⟨html⟩\⟨body⟩\⟨h1⟩	3
p_4	Document\⟨html⟩\⟨body⟩\⟨br⟩	3
p_5	Document\⟨html⟩\⟨body⟩\List	3
p_6	Document\⟨html⟩\⟨body⟩\⟨h1⟩\Tech	1
p_7	Document\⟨html⟩\⟨body⟩\⟨h1⟩\World	1
p_8	Document\⟨html⟩\⟨body⟩\⟨h1⟩\Local	1

Fig.2 Paths of tokens and their support

ii) Matrix representation of clustering

1. In this we are defining the clustering of web documents.

2. Here *cluster* ci is denoted by a pair of ***(Ti,Di)*** where Ti is the set of paths representing the templates of ci and Di is a set of documents belonging to ci.

3. Here we are finding out the essential path matrix through clustering.

MT denotes the information of each cluster with its templates

MD denotes the information of each cluster with its member documents.

Cluster C={c1,c2} ,where
c1=({p1,p2,p3,p4,p5},{d1,d2,d3})
c2=({p1,p2},{d4})

Minimum description length principle

1. In order to manage the unknown number of cluster and to select good partitioning from all possible partitions of HTML documents.

2. The MDL principle states that the best model inferred from a given set of data is the one which minimizes the sum

The length of the model in bits.

- The length of the coding of the data.

We refer the above sum of the model as MDL cost of the model

3. In this we calculating the cost and then finding out the best cluster.

H(X) =Σ-Pr(x) log2Pr(x) and

L (M) =|M|.H(X**)**

iii) Implementation steps

1. Selected no. of websites and web pages.

2. Extract web pages from different no. of web sites.

3. Each and every web site can provide different templates.

4. Label based training data and cluster the web pages.

5. It form clusters with different number heterogeneous templates.

6. It give the solution as a best solution and best cluster.

7. Best cluster can be providing as an optimal cluster.

IV.. SYSTEM DESIGN

System is developed into subsystems. The sub-project is divided into sub-tasks:

1. Develop a model with high-level system specification and verify it.

2. Developing a systematic method to refine the specification into synthesizable code and a prototype tool.

Today it is common to specify systems on higher levels using some natural language (e.g. English). Large amounts of information must be handled, problems arise. Errors are detected late in the design cycle .

By making the initial system specifications in a formal language at a high abstraction level verified/tested. Ambiguities and inconsistencies can be avoided, errors can be discovered earlier, and the design iteration cycles can be shortened, thereby reducing development times.

Further, most of the languages that are used for implementation of HW / SW designs (e.g. VHDL, C++) do not lend themselves well to formal verification. A lack of formal semantics sometimes causes ambiguities in the interpretation of the designs.

Our goal is to develop functional system specification method for telecom systems . The specification language in which the system level functions will be developed will have a formal semantics in order to support formal verification of specifications

V.SOFTWARE TESTING

Testing[5] is an important phase in the development life cycle of the product. It is used to detect the errors. It perform a very critical role for quality assurance and ensuring the reliability of the software.The testing determines the program reliability of the software. During the testing, the program to be tested and executed with a set of test cases and the output of the program for the test cases was evaluated to determine whether the program is performing as expected. Thus, a series of testing was performed on the system before it was ready for implementation. Testing is process of technical investigation, ie intended to reveal the quality-related information about the product with respect to context in which it is intended to operate.It is not limited to the process of executing a program or application with the intent of finding errors.

Unit testing, validates the internal program logic functionality, whether the program input produces valid outputs or not. In any application, individual software units are tested before the integration. This is called structural testing; that relies on knowledge of its construction. Unit tests perform at component level , business application, and/or system configuration level. Unit tests ensure that it gives the documented specifications, defined inputs and expected results.

The integrated software components are actually running as one program or not. Even though all components are individually working satisfactorily and combining these components giving correct and consistent results are not done by integration testing. Integration testing is aimed to expose the problems, when combining the components.

The functions are tested and availability of functions as specified by the business, technical requirements, and user manuals. Functional testing concerned on the valid Input, invalid Input, functions, output and interfacing Procedures. Functional tests focused on requirements, key functions, or special test cases and systematic coverage.

System test is based on process descriptions, flows, emphasizing pre-driven process links and integration points.

White Box testing is used to test areas that cannot be reached from a black box level. There is no knowledge on the inner workings.

Black Box Testing:

We cannot see into it, provides inputs and responds to outputs without considering how the software works.

VI.CONCLUSION AND FUTURE SCOPE

It uses novel approach of the template detection from heterogeneous web documents. Matrix form uses to represent the documents and path, and matrix uses the MDL principle to manage the unknown number of clusters and to select good partitioning from all possible partitions of documents, and then introduced our extended min hash technique to speed up the clustering process.

In real life data sets confirmed the effectiveness of our algorithms from this approach we can reduce the cost, and also we can reduce burden of machines. There are many approaches are implementing to increase performance, speed up the process.

VII. **BIBLIOGRAPHY**

1. The Internet Book: Everything You Need to Know about Computer Networking and How the Internet

Works Douglas Comer, 4th Edition, Prentice Hall,2007.

2. Internet Books for Educators, Parents, and Students, Jean Reese, Libraries Unlimited, 1999.
3. Blanchard, B. S., & Fabrycky, W. J.(2006) *Systems engineering and analysis* (4th ed.) New Jersey: Prentice Hall.
4. Computer World, 2002, Retrieved on June 22, 2006 from the World Wide Web.
5. Gelperin, D.; B. Hetzel (1988). "The Growth of Software Testing". *CACM* **31** (6). ISSN 0001-0782.
6. Hash Functions". *cse.yorku.ca*. September 22, 2003. Retrieved November 1, 2012. "the djb2 algorithm (k=33) was first reported by dan bernstein many years ago in comp.lang.c.
7. A. Tanay. R. Sharan, and R. Shamir, "Biclustering Algorithms: A Survey", In *Handbook of Computational Molecular Biology,* Edited by Srinivas Aluru, Chapman (2004).
8. Koch, Peter-Paul (May 14, 2001). "The Document Object Model: an Introduction". *Digital Web Magazine*. Retrieved January 10, 2009.